Tree Genetics & Genomes (2012) 8:1359–1369
DOI 10.1007/s11295-012-0523-6

ORIGINAL PAPER

A first insight into peach [*Prunus persica* (L.) Batsch] SNP variability

María José Aranzana · Eudald Illa · Werner Howad · Pere Arús

Received: 1 December 2011 / Revised: 13 April 2012 / Accepted: 9 May 2012 / Published online: 10 June 2012

Abstract Three factors may have reduced the diversity at both individual gene and whole genome levels in cultivated peach: its self-compatible mating system, the narrow genetic basis of most commercial cultivars, and the recent strong selection towards agronomically interesting traits. Previous diversity analyses with markers such as simple sequence repeats (SSRs) have revealed low levels of genetic variability. Here, we sequenced 23 genome-wide distributed DNA fragments in 47 occidental peach varieties, also observing reduced variability levels. On average, there was one single nucleotide polymorphism (SNP) every 598 bp and one indel every 4,189 bp. As expected, variability was higher in non-coding than in coding regions (one SNP every 390 non-coding bp versus one in 1,850 bp in coding DNA). In general, SNPs were observed at relatively high frequency, mean minor allele frequency=0.225, meaning that a large proportion of the SNPs discovered by sequencing similar germplasm will be useful for other purposes, such as association mapping. The average heterozygosity of the varieties was 0.28, with a low correlation between SSR and SNP heterozygosity. The whole sequence of two candidate genes, a pectate lyase 1 candidate for fruit firmness (CGPAA2668) and a sucrose synthase 1 candidate for sugar content (CGPPB6189), in the 47 varieties revealed that they both may have suffered a process of balancing selection.

Communicated by A. Abbott

Electronic supplementary material The online version of this article (doi:10.1007/s11295-012-0523-6) contains supplementary material, which is available to authorized users.

M. J. Aranzana · E. Illa · W. Howad · P. Arús (✉)
IRTA, Centre de Recerca en Agrigenòmica CSIC-IRTA-UAB-UB, Campus, UAB,
Bellaterra (Cerdanyola del Vallès),
08193 Barcelona, Spain
e-mail: pere.arus@irta.cat

Keywords *Prunus* · Cultivars · Breeding · Crop evolution · Genetic diversity

Introduction

Molecular marker variability, using isozyme genes (Byrne 1990) and simple sequence repeat (SSR) markers (Mnejja et al. 2010), has shown that peach is the least genetically variable of the *Prunus* crops, that also include apricot, cherry, Japanese plum, and almond. The fact that the gametophytic self-incompatibility system is not operative in peach but functional in the other species results in a high level of selfing (Miller et al. 1989). Homozygosity is a consequence of selfing which, when coupled with selection for different agronomic characters and for progeny phenotypic uniformity, leads to erosion of the genetic variability. In addition, the cultivars currently commercialized in Europe and America come from a very limited gene pool, used by the initial US breeders about one century ago (Scorza et al. 1985), resulting in a bottleneck that further diminished the level of variability.

A large set of peach cultivars from Europe and North-America has been analyzed with SSRs by Aranzana et al. (2003a; 2010). Despite having a level of variability sufficiently high for the individual identification of virtually all cultivars, these SSRs were found to be relatively less variable than in other species. The collection of cultivars studied was structured in subpopulations, generally corresponding to certain key commercial characters: peaches, nectarines and non-melting flesh (canning) peaches. High conservation of linkage disequilibrium has also been detected with a collection of 50 SSRs (Aranzana et al. 2010), as expected considering the bottleneck that occurred at the beginning of modern peach breeding.

DNA sequence variability was studied in a collection of 47 cultivars selected to be representative of the variability of the species on the basis of SSR variability using a set of 23 peach DNA sequences, RFLP genomic probes, and ESTs, of known position on the map (Dirlewanger et al. 2004; Illa et al. 2011) and genome (http://www.rosaceae.org/). Two of the EST sequences, corresponding to a pectate lyase and a sucrose synthase gene identified by Illa et al. (2011) as candidate genes for fruit texture and fruit glucose content, were studied at the whole sequence level. These results provide a first insight into the sequence variability of peach and allow us to study the variability of haplotypes in this species where high linkage disequilibrium (LD) conservation is expected.

Material and methods

Plant material

To evaluate the levels of sequence variability in commercial peaches, 47 peach varieties were selected from a collection of 224 previously analyzed with 50 SSR markers (Table 1). These varieties have been shown to be genetically distant and representative of different subpopulations (Aranzana et al. 2010). Genomic DNA was isolated from young leaves as previously described by Viruel et al. (1995).

DNA sequencing

Genome-wide sequence variability

Among the sequences available in *Prunus* at the Genome database for Rosaceae (http://www.rosaceae.org/), we selected 40 regions sequenced in peach, evenly distributed along the *Prunus* reference map. Ten of them derived from RFLP genomic probes and the rest (30) from ESTs. Specific primer pairs (Table 2) were designed for each region using the Primer3 software (Rozen and Skaletsky 2000; http://frodo.wi.mit.edu/) to amplify fragments of about 450 bp, avoiding amplification of SSR regions.

The primers were first tested in two peach varieties, “Alexandra” and “Calante”, identified as high and low heterozygous, respectively, with SSRs (Aranzana et al. 2010). For sequencing, 40 ng of peach genomic DNA were first amplified in a total volume of 20 µl with 1XPCR buffer, 1.5 mM $MgCl_2$, 0.5 mM dNTPs, 0.25 µM of each primer, and 1.5 U of GoTaq® (Promega) using the following conditions: 2 min at 94°C; 35 cycles of 15 s at 94°C; 1 min at the appropriate annealing temperature; and 1 min at 72°C, followed by a final extension step of 5 min at 72°C. PCR products were purified using Sephadex™ G-50 (GE Healthcare Life Science) as described by Till et al. (2006). DNA quantity was measured using a spectrophotometer (NanoDrop technologies, Wilmington, DE, USA) and confirmed by electrophoresis on 1 % TBE agarose gel. Forward primers were used for sequencing the fragments using the BigDye® Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA), according to the manufacturer's protocol, in an ABI Prism® 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Sequences were visualized and manually edited with Sequencher 4.8 software (Gene Codes Corporation; Ann Arbor, MI, USA). Fragment ends were trimmed to remove low-quality sequence. Among the analyzed sequences, the 23 yielding high quality, unique sequences were selected (Table 2) and sequenced in 47 varieties.

To identify the coding and non-coding regions, RFLP sequences were blasted against the *Populus* genome database (http://www.populus.db.umu.se/) and the NCBI site (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Amplified EST sequences were also aligned against the FASTA sequences from which the primers were designed to detect intronic regions.

Sequence variability of two candidate genes

Two of the polymorphic fragments corresponded to the CGPAA2668 (pectate lyase 1, candidate for fruit firmness) and CGPPB6189 (sucrose synthase 1, candidate for sugar content) candidate genes (Illa et al. 2011) and were selected for whole gene sequencing. To amplify both genes, primers were designed by blasting the candidate gene fragments with the ESTs available in the GDR database (Jung et al. 2008). CGPAA2668 and CGPPB6189 were fully amplified in the 47 varieties with three and four primer pairs, respectively. The resulting amplified fragments were sequenced with four and seven primers (Table 3). Results were aligned with Sequencher 4.8 software (using the large gap algorithm) with additional manual adjustments in the case of long insertion/deletion (indel) polymorphisms.

Sequence variability analysis

For each single nucleotide polymorphism (SNP), allelic and genotypic frequencies and observed and expected heterozygosity (Ho and He, respectively) were calculated and deviation from the Hardy–Weinberg equilibrium (HWE) was tested. Two of the polymorphic fragments detected contained more than one SNP in heterozygosis. In each fragment, the SNPs were linked in a whole haplotype and consequently there was no phase ambiguity. Ho was calculated for each cultivar. Additionally, for each polymorphic fragment, we calculated two estimates of nucleotide polymorphism by quantifying the number of segregating sites, θ_W (Watterson 1975), and the nucleotide diversity, π, i.e.,

Table 1 Characteristics of the 47 peach cultivars used

Cultivar	Pedigree	Breeder, Country	Fruit traits	SSR Ho	SNP Ho
Admiral Dewey	Alexander op.	–	PYM	0.00	0.00
Alexandra	Fayette×Royal Gold	Zaiger's Genetics (Zaiger), USA	PWM	0.50	0.60
Aline	O'Henry×Giant Babcock	Zaiger, USA	PWM	0.71	1.00
Andross	Dix 5A-1×Fortuna	UCD, USA	PYN	0.00	0.33
Babygold-7	(Lemon Free×PI35201)×NJ196	RU-NJ, USA	PYN	0.43	0.80
Big Top	–	Zaiger, USA	NYM	0.43	0.67
Binaced	–	Servicio de Investigación Agraria, Zaragoza, Spain	PWN	0.00	0.17
Calabacero	–	TC, Spain	PYN	0.14	0.20
Calante	–	TC, Spain	PYN	0.00	0.00
Catherina	NJC95×D42-13w	RU-NJ, USA	PYN	0.67	1.00
Chinese Cling	–	Imported from China	PWN	0.29	1.00
Cresthaven	Kalhaven×SH309	Michigan State University, East Lansing (MSU), USA	PYM	0.00	0.20
Dellys	–	Escande, France	NWM	0.29	0.17
Early Crawford	–	Crawford (1841)	PYM	0.14	0.33
Early Elberta	Elberta op.		PWM	0.71	0.67
Elberta	Chinese Cling op.(perhaps×Early Crawford)	S. H. Rumph, USA	PYM	0.67	0.83
Escarolita	–	Traditional cultivar, Spain	PYN	0.14	0.00
Fay Elberta	Elberta op.		PYM	0.29	0.50
Festina	–	Escande, France	NWM	0.00	0.17
Flavor Gold	Rhone Gold×Royal Gold	Zaiger, USA	NYM	0.14	0.67
Flavortop	Fairtime op.	Fresno, USA	NYM	0.29	0.50
Flavour Giant	–	Zaiger, USA	NWM	0.14	0.17
J.H.Hale	Chance sdlg, possibly of Elberta	J. H. Hale	NYM	0.40	0.60
Jesca	Sel. from Amarillos Tardios de Calanda	TC, Spain	PYN	0.29	0.33
Large White	–	U.S.D.A. Baton Rouge, Lousiana, USA	PWM	0.14	0.67
Maria Bianca	Honey Dew Hale×Michelini	DOFI, Italy	PWM	0.29	0.50
Michelini	–	Michelini, Savona, Italy	PWM	0.43	0.50
Nectaross	Stark Redgold×Le Grand	ISF, Rome (ISF), Italy	NYM	0.29	0.50
Paraguayo Delfin	–	Traditional cultivar, Spain	FWM	0.29	0.50
Queen Crest	Maycrest mutation	Balakian Reedley, California, USA	PYM	0.13	0.50
Queen Giant	–	Zaiger, USA	NWM	0.14	0.67
Redhaven	Halehaven×Kalhaven	MSU, USA	PYM	0.40	0.50
Redwing	Babcock×Stensgaard July Elberta	Armstrong, USA	PWM	0.71	0.83
Rio Oso Gem	Late Crawford op.	–	PYM	0.14	0.00
Royal Glory	May Grand op.	Zaiger, USA	PYM	0.67	0.50
Royal Moon	–	Zaiger, USA	PYM	0.29	0.67
Royal Prince	–	Zaiger, USA	PYM	0.29	0.67
Seduction	–	Maillard, France	PYM	0.17	0.17
Silver Gem	May Grand×Chance Seedling	Zaiger, USA	NWM	0.57	0.50
Snow Queen	–	Armstrong, USA	NWM	0.43	0.25
Starlite	FV89-14×Springtime	Byron, USA	PWM	0.14	0.17
Summer Grand	Late Le Grand×Early Sun Grand	Bradford, USA	NYM	0.67	0.80
Suncrest	Alamar×Gold Dust	Fresno, USA	PYM	0.14	0.40
Super Crimson Gold	Zee Gold×Early Sun Grand	Zaiger, USA	NYM	0.33	0.50
Tendresse	–	Maillard, France	PWM	0.71	0.83
Villa Giulia	Catherina op.	ISF, Italy	PYN	0.14	0.33
Voluptia	–	ISF, Italy	PWM	0.43	0.50

P peach, *N* nectarine, *W* white flesh, *Y* yellow flesh, *M* melting flesh, *N* non-melting flesh, *Ho* observed heterozygosity

Table 2 Description of the RFLP genomic probes and ESTs sequenced in 47 peach cultivars

Name	GeneBank accession	C	Map position (cM)[a]	Peach sequence position	Primer F	Primer R
CGPPC7442	AF367442	1	1:14	2804161..2804764	GCATGGAAAGAAAAGATGTC	CGTTCAATCTCTCTGAAACC
AG116	BH023813	1	4	10175205..10175634	TGATGTCAATCAGAGTGTGTGG	TCCTATGAGGCTCTCGAACA
PP_LEa0005O07f	BU040407	1	66	10746023..10746892	ATGGGGAAATATGAATGGTCAG	CTGGTTTGAGTACCAGGAAAGG
AG105	BH023887	1	48	33128052..33128478	GCTGTTGCAGCAAACAAAAG	CACACCATGCAATGTATCACG
AG35	BH023841	2	25	17821458..17822016	GGACAAGGTTTACATTGGCTGT	GCAGCCTGCATTTGTTTAGTTT
PP_LEa0017B03f	BU043752	2	45	21130516..21132396	CACACAGAGATGATGATGCAGA	GACCCAAATGCTATAGCTGCAC
CGPPC9097	DW359097	2	2:34	21733970..21734315	TCTAACGTTGGAGGAGAAGA	TCAGGAGGAGTTTTGAGAGA
CGPPC8197	DY638197	2	2:38	22456186..22456664	CACTCATTCTTGTTGGAGGT	GTGAGCATGTGATCTTGATG
CGPPC7895	DY637895	3	3:06	96450..96746	AGTGTATGGTCCAGTCAAGG	CCATCTTCTACTGCTGGAAC
CGPPC4457	DY634457	3	3:49	19173416..19173706	GTGACACCAAGAAGATGGAT	TTTCAACTTGTTGCTCCTTT
PP_LEa0003O13f	BU039800	4	5	scaffold_10:289314..289959	GAATCTCCCTCTCTCTGTCTGC	AGAACACGGTTCCAAATCCTAA
AG62	BH023808	4	28	8692980..8693382	TTCAATTATAAGCGCTTTTTCCA	TTTGAAAATTGCAAGATGAAAAT
PP_LEa0001F16f	BU039139	5	33	3612802..3613419	GAGACTTTTGGAAGGGAAGGAT	CGCTCGATTTTTGGTCTTCTTA
AG114	BH023921	5	15	8944055..8944471	CTACCTTATCGCGGCTTCTTC	GCTCCTGGATGGTATATTTGCT
CGPAA2668	CB822668	5	5:41	14194017..14194620	AAGCTTGAAGCCCTTGCATA	TCACTTTGGAGAAGGGTTGG
CGPPC2807	AJ872807	5	5:41	15127115..15128665	GGATCTGGGTTTGTATTGTT	ACAAGCAGAAATTGGATGAT
CGPPC6491	BU044123	6	6:65	21764941..21766776	TTGTGTTGTAGGTGGTGAAA	TCTGGGGATACAGTAAGAAAA
PP_LEa0029C22f	BU047210	6	74	25072623..25073400	CAAGTTCCTCAAATCAAGTGCT	TAACTTAACGATTCCGTCGTGA
CGPPC7441	DY647741	6	6:84	27256945..27258575	GCAACGTTCAGTACGCCTTT	CTGCACTGTTGCAATGCTCT
AG101	BH023905	7	10	3515203..3515614	CGCCATGTTCTTTTTCTTCAAT	TAATGGATCCAGATCCTCTGCT
CGPPC3657	DY633657	7	7:31	12700590..12701833	CCTGCCCGTCTAATGATAGT	TGTTGGTGACCTACCAGTTG
CGPPB6189	AJ876189	7	7:56	18757056..18757387	ACTCCGGTGAGAGTCAAGAG	ACCTGTGGACTTCCAACATT
AG112	BH023902	8	13	3583449..3583858	TAATTTGGCATCCATTGCATTA	CTAAAAGGCAAAAATGGGCAAT

C chromosome (macromolecule in the case of whole sequence data)

[a] Map position terminology is as in Dirlewanger et al. (2004) for markers mapped in the whole TxE population or as in Howad et al. (2005) for those bin mapped

Table 3 Primer pairs used for amplification pectate lyase 1 (*PpPL1*; 2,239 bp of consensus sequence) and sucrose synthase 1 (*PpSUS*; 3,941 bp)

Primer name	Sequence	Primer name	Sequence
PpPL1_0033_F	TCTACCATTATTCAAGGCTTGCT	PpPL1_0748_R[a]	CAATGGCATTTTTCCCAAAC
PpPL1_0409_F[a]	CAAAGGTAAGGCTCCACCAA	PpPL1_1745_R[a]	AAACCGAACCCATCATAAAAT
Pp_PL_1398_F[a]	GGTCCATGATTTTGATGTTGC	PpPL1_2392_R	AGAGAAGGAGGAAAGACAGAGC
PpSUS_0032_F[a]	TGGAATTATTGACTTGGTGGTG	PpSUS_1083_R	TGAGAGAAATCGATCTGTATAAGGAA
PpSUS_0732_F[a]	GAGGAACTTGTTGATGGAAGG	PpSUS_2109_R[a]	CTTGGAAAGTTCCTGGATCG
PpSUS_1754_F[a]	ACCTCATTGTGATGGCTTGA	PpSUS_3583_R[a]	GAAGCCATAAACTCCGGTGA
PpSUS_3245_F[a]	CTGTGGACTTCCAACATTCG	PpSUS_4067_R	GTGCGTTCAACAAAAAGCAA
PpSUS_3117R[b]	TCTGAACTGCCCATTCAACTT		

[a] Primers used to obtain the whole gene sequence

[b] Primer used only to complete the gene sequence

the mean percentage of nucleotide differences among all pairwise comparisons (Nei 1987). To allow comparison between different regions, we estimated these parameters for each site. Neutrality of the mutations was tested through Tajima's *D* statistic (Tajima 1989). These parameters were calculated with the software DnaSP v5 (Librado and Rozas 2009).

Variability comparison between SSRs and SNPs

To compare the variability detected by SSR and SNP markers, we selected the six SSRs closest to the fragments found to be polymorphic (Supplementary data, Table S1) as described in Aranzana et al. (2010).

HWE deviation of the SSR and SNP markers in the 47 cultivars was analyzed with GDA software (Lewis and Zaykin 2001). Additionally, two genetic distance matrices were constructed with the NTSYSpc v 2.10t program (Rohlf 1994) with SSR and SNP data for all the analyzed cultivars as described in Aranzana et al. (2010). Both matrices were compared through a two-way Mantel test with the MxComp procedure of the NTSYSpc V. 2.10t program.

The correlation between the heterozygosity levels detected with both types of markers was calculated with the JMP software package version 8.0.1 (SAS Institute Inc, Cary, NC) by the REML method.

Results and discussion

Sequence variability

In total, we sequenced 23 DNA regions in 47 cultivars, obtaining 8,379 bp/cultivar (i.e., 393,813 bp sequenced as a whole), 4,677 bp corresponding to non-coding regions and 3,702 bp to coding regions (Table 4). Nucleotide variation was observed in seven out of the 23 sequenced fragments (30 %), with 14 SNPs and two indels, corresponding to one SNP every 598 bp and one indel every 4,189 bp. As expected, variability in non-coding regions was higher than in coding regions (Ching et al. 2002; Lijavetzky et al. 2007; Micheletti et al. 2011), with one SNP every 390 non-coding bp versus one in 1,850 bp in coding DNA. According to these data, the proportion of fragments found with at least one polymorphism is much lower than that in other species also subjected to bottlenecks and strong selection, such as sugarcane where sequencing projects have found 86–94 % of the fragments (depending on the sample set) to be polymorphic (Bundock et al. 2009). Similarly, SNPs were observed at a lower density compared with other crops such as melon, tomato, grape, maize, or apple (Table 5). The observed low levels of sequence variability are consistent with those obtained using molecular markers such as AFLPs and SSRs (Aranzana et al. 2003b, 2010) in peach and with isozymes (Byrne 1990) and SSRs (Mnejja et al. 2010) in other *Prunus* species. Direct sequencing of genomic fragments (usually ESTs) as a tool for SNP discovery has been successfully used in different plant species; however, the low number of polymorphic fragments and SNP density found here implies that this method may be less efficient in peach, supporting the need for high-throughput sequencing strategies for this purpose.

All of the SNPs were found to be biallelic, 64 % due to transitions and 36 % to transversions. Although, probabilistically, the expected proportion between transitions and transversions is 1:2, a bias towards transitions is frequently observed, probably as a consequence of greater purifying selection against transversions (Keller et al. 2007) that may vary for different organisms (Strandberg and Salter 2004). The transition/transversion ratio observed here (1.77) is similar to that observed in grape (1.56 by Salmaso et al. 2004 and 1.46 by Lijavetzky et al. 2007) and potato (1.5 by Simko et al. 2006) and higher than that observed in apple (1.27 by Micheletti et al. 2011).

Table 4 Polymorphism of SNPs and indels found in a set of 47 peach cultivars

Fragment	No. of bp	Total no. of SNPs	Non-coding			Coding		
			No. of bp	No. of SNPs	No. of indels	No. of bp	No. of SNPs	No. of indels
CGPPC7442	518	0	369	0	0	149	0	0
AG116	318	0	195	0	0	123	0	0
PP_LEa0005O07f	594	0	120	0	0	474	0	0
AG105	318	1	318	1	0	0	0	0
AG35	424	0	213	0	0	211	0	0
PP_LEa0017B03f	497	0	104	0	0	393	0	0
CGPPC9097	271	0	144	0	0	127	0	0
CGPPC8197	395	0	253	0	0	142	0	0
CGPPC7895	202	0	128	0	0	74	0	0
CGPPC4457	228	0	110	0	0	118	0	0
PP_LEa0003O13f	520	0	0	0	0	520	0	0
AG62	323	0	261	0	0	62	0	0
PP_LEa0001F16f	490	0	490	0	0	0	0	0
AG114	329	0	123	0	0	206	0	0
CGPAA2668	566	1	288	1	0	278	0	0
CGPPC2807	392	1	274	1	0	118	0	0
CGPPC6491	214	0	214	0	0	0	0	0
29C22	630	0	192	0	0	438	0	0
CGPPC7741	660	1	537	1	0	123	0	0
AG101	278	1	278	1	0	0	0	0
CGPPC3657	187	0	90	0	0	97	0	0
CGPPB6189	281	5	83	3	0	198	2	0
AG112	262	4	262	4	2	0	0	0
Total	8,379	14	4,677	12	2	3,702	2	0

The number of polymorphic sites in polymorphic loci (including SNPs and indels) varied from one to six, with an average of 2.3 (Table 4). Five out of the seven polymorphic fragments contained only one polymorphism, all of them in non-coding DNA. In contrast, the two remaining loci (CGPPB6189 and AG112) were highly polymorphic, the former with five SNPs, two of them in coding DNA, and the latter with four SNPs and two indels (of 1 bp and 2 bp), all occurring in non-coding DNA. As sequences were obtained from PCR-amplified genomic DNA, each sequence contained the two DNA strands. This can produce phase ambiguity in the case of multiple polymorphisms per fragment, and if large indels occur in heterozygosis, base calling becomes unfeasible. However, in both loci, the homozygous genotypes showed that all SNPs were linked, yielding two haplotypes per fragment and, consequently, three genotypes, leaving no ambiguity for phase determination. Moreover, in the AG112 fragment, the cultivars carrying the less frequent haplotype in homozygosis were also homozygous for the two indels, suggesting that they were linked to the SNPs, so we can assume that the heterozygous cultivars for the SNP variants were also heterozygous for the same two indels observed in the homozygous genotypes.

Genetic variability, measured as θ_w, gave values ranging from 0.0003 to 0.0035 with an average of 0.00129 (Table 6, Fig. 1). Nucleotide diversity, π, is a parameter that depends on the number of SNPs as θ_w, and also on their frequency. These values were low, ranging from 1.7×10^{-4} (in CGPPC2807) to 6.8×10^{-3} (in CGPPB6189) with an average of 2.1×10^{-3} , i.e., we expect two randomly chosen sequences of 1,000 bp selected from one of the polymorphic fragments to differ, on average, in about two sites.

Observed mean θ_w values were similar to those reported for soybean (0.00097; Zhu et al. 2003) and about 3.5 and 7.5-fold lower than that observed in grapevine (0.0046 Lijavetzky et al. 2007) and maize (0.0096; Ching et al. 2002). When taking into account allele frequencies, there was more similarity in variability levels (measured as π), with peach being a little over 1.5 times higher than soybean (π = 0.0012) and just two and three times lower than grapevine and maize (π = 0.0051 and 0.0063, respectively). This enhancement is a consequence of the relatively high allele frequencies observed. Here, SNP minor allele frequencies (MAF) ranged from 0.036 (in CGPPC2807) to 0.420 (in CGPPC7741) with a mean value of 0.225 (when

Table 5 Comparison of SNP variability in various plant species

Species	No. of bp/SNP	No. of bp/indel	No. of genotypes	Length DNA sequenced per genotype (kb)	No. of bp/SNP (coding)	No. of bp/SNP (non-coding)	% coding sequence	Reference
Apple	52	333	135	3.4 or 11.3	48	40	46	Micheletti et al. (2011)
Maize elite inbred lines	61	126	36	6.9	124	31	34	Ching et al. (2002)
Grapevine	64	1,932	10	100.5	69	47	81	Lijavetzki et al (2007)
Cultivated tomato	150	680	31	23.1	NA	NA	NA	Labate et al. (2009)
Melon	441	1,666	2	15.0	NA	NA	NA	Morales et al. (2004)
Peach	598	4,189	47	8.4	1,850	390	44	This paper

NA No data available

considering only unlinked SNPs), and 64 % of them had a frequency higher than 0.2, while for example, in grapevine, 50 % of alleles had MAF>0.2 (Lijavetzky et al. 2007) (Fig. 2). This contrasts with allele frequencies observed in species more variable than peach, such as apple, with 26–42 % of alleles with MAF>0.2 after re-sequencing two different sets of *M. x domestica* germplasm (Micheletti et al. 2011). Our results suggest that a large proportion of the SNPs discovered on sequencing peach occidental germplasm will be useful for association mapping purposes, where MAF is usually set at ≥5 %, as well as for inclusion in large-scale genotyping platforms, where only robust SNPs are desired.

Tajima's *D* statistics, which detects departure from neutrality of mutations by comparing θ_w and π estimates, gave values ranging between −0.778 and 2.078, with a mean of 1.06. Under neutral equilibrium, Tajima's *D* is expected to be zero. Significant departure from neutrality ($p \leq 0.05$) was only observed in one of the fragments, CGPPB6189, with 2.078. This fragment amplifies part of a candidate gene that encodes a sucrose synthase. Positive Tajima's *D* values can indicate balancing selection, which tends to maintain several alleles at intermediate frequency (Wright and Gaut 2005). However, SSR data in the region around this locus (linkage group 7, 7:56 bin) do not show an increase of heterozygosity compared to other genomic regions (data not shown).

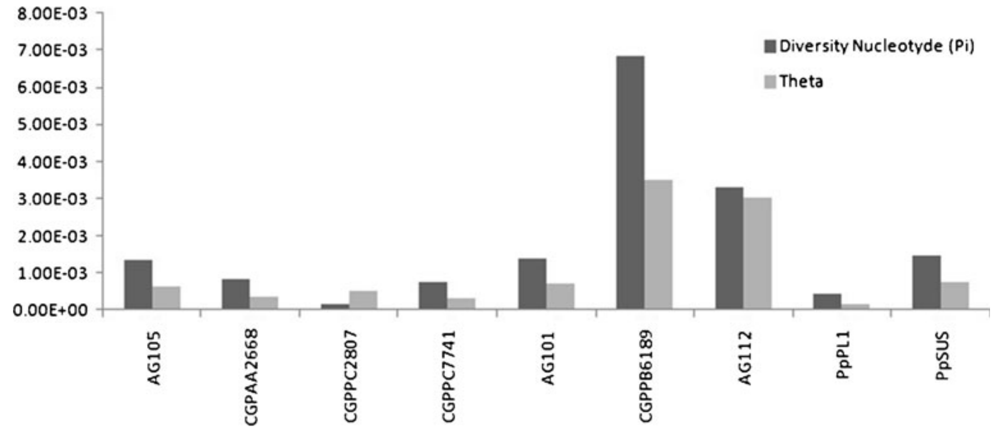
At each of the seven polymorphic loci, only two alleles (haplotypes) were amplified. Ho ranged from 0.023 to 0.477 (mean Ho=0.263) and He from 0.022 to 0.487 (mean He=0.307). These values are lower than those observed with SSR markers in commercial peach varieties, where Ho and He have been estimated to be 0.35 and 0.46, respectively (Aranzana et al. 2010), such that single SSRs are more informative than single SNPs for variability studies. This has been observed elsewhere. For example, Laval et al. (2002) calculated that k-1 times more biallelic markers are needed to achieve the same genetic distance accuracy as a set of microsatellites with k alleles. In peach, the average number of alleles per SSR ranges from about 3.5 to 7.3, depending on the cultivars and SSRs used (Wünsch et al. 2006; Dirlewanger et al. 2002; Testolin et al. 2000; Sosinski et al. 2000; Aranzana et al. 2003a, 2010). This means that, to get the same accuracy as with 100 SSR markers, we would need between 250 and 630 SNPs.

With population admixture, genotypic frequencies may deviate from those expected under panmixia. Testing for HWE as a measure of population admixture may predict false positives in association studies (Deng et al. 2001; Tiret and Cambien 1995). All SNPs found here were in HWE (considering $p \leq 0.05$). This contrasts with the generalized departure from HWE previously detected with SSRs in peach (Aranzana et al. 2003a). HWE equilibrium departures can be caused by intrinsic factors in the studied sample, such as population admixture and selection, but also by specific marker characteristics such as mutation rates (Deng et al. 2001). In the case of selection, HWE departures will not only affect the marker but also a relatively large region around the genomic region analyzed. For a more realistic comparison of SNPs and SSRs, we selected a set of six SSRs from those analyzed by

Table 6 Variability parameters of seven polymorphic DNA sequences in peach

	AG105	CGPAA2668	CGPPC2807	CGPPC7741	AG101	CGPPB6189	AG112	Avg.	Max	Min
No. of SNPs	1	1	1	1	1	5	4			
Sample size	46	47	43	45	45	47	46			
Sequence length	318	566	392	660	278	281	262			
π	1.35E-03	8.50E-04	1.70E-04	7.50E-04	1.38E-03	6.84E-03	3.29E-03	2.09E-03	6.84E-03	1.70E-04
θ	6.20E-04	3.50E-04	5.10E-04	3.00E-04	7.10E-04	3.48E-03	3.03E-03	1.29E-03	3.48E-03	3.00E-04
Tajima's *D*	1.37683	1.7046	−0.7779	1.76	1.11462	2.07845	0.16937	1.061	2.078	−0.778

Fig. 1 Estimates of variability (π and θ) for the polymophic fragments and the candidate genes *PpPL1* and *PpSUS* fully sequenced



Aranzana et al. (2010), adjacent to the polymorphic fragments, and reanalyzed them in the same set of varieties. Three of the SSRs departed from HWE ($p \leq 0.05$): BPPCT020 about 153 kbp from AG105, BPPCT038 464 kb and 469 kb from CGPAA2668 and CGPPC2807, respectively, and UDP96-008 2.2 Mbp from CGPPC4457. LD in peach has been estimated to extend 13–15 cM (Aranzana et al. 2010). Considering a rough correspondence of 430 kbp/cM, LD extends about 5.59–6.45 Mbp, so we can consider that the analyzed SSRs and SNPs are linked. This means that the departure from HWE is more probably due to their different mutational properties.

The average heterozygosity of the cultivars was 0.28, ranging from 0 to 0.71. The most heterozygous were “Aline”, “Early Elberta”, “Redwing” and “Tendresse”, whereas “Admiral Dewey”, “Andross”, “Binaced”, “Calante”, and “Festina” were homozygous for all of the sequenced fragments (Table 1). “Admiral Dewey” and “Calante” were also homozygous at the six SSRs tested, as were “Escarolita” and “Rio Oso Gem”. In contrast, “Aline”, “Elberta”, and “Chinese Cling” were heterozygous at all six SSRs. The correlation of the heterozygosity found between SSR and SNP data was low ($r=0.504$). To compare SSR and SNP variability data, a distance matrix was also constructed for both SNP and SSR data. On comparing both matrices through a Mantel test, no correlation between them was observed ($r=-0.057$). The reason could be that SNP-based distances are due almost entirely to drift, while SSR-based distances are also due in part to mutation (Hamblin et al. 2007). The low correspondence between both types of information is shown graphically in Fig. 3, where an SNP matrix alignment is plotted against the SSR distance tree.

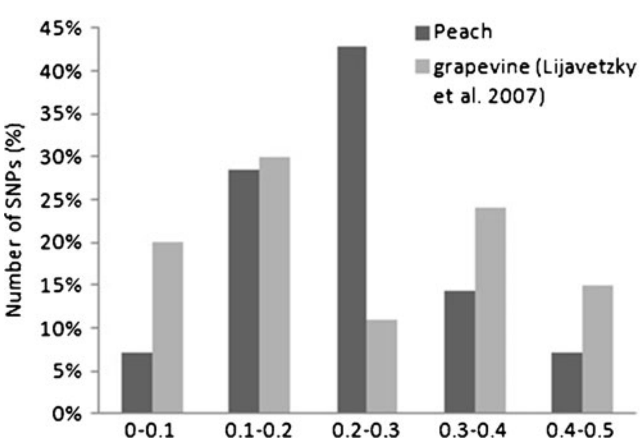


Fig. 2 Comparison of SNP allele frequency distribution in peach and grapevine

Up to now, most peach variability has been assessed with SSRs. This information is now being used to select varieties to be included in sequencing projects, such as those oriented to SNP discovery. Our results suggest that the selected varieties may not fulfill the expectations concerning variability and heterozygosity that SSRs predict.

Variability at two candidate genes

Among the seven polymorphic fragments found, two were candidate genes for important economical characters and the whole gene was sequenced in the 47 cultivars (Fig. 4). One (CGPAA2668, accession number CB822668) corresponded to a region of a pectate lyase, a gene involved in cell wall degradation and fruit softening which is consequently considered to have a role in ripening (Marin-Rodriguez et al. 2002). A quantitative trait locus (QTL) for flesh firmness has been detected in the region of linkage group 5 that contains this gene in an F2 population between the two peach cultivars Ferjalou Jalousia® and Fantasia (E. Dirlewanger, pers. comm.). The other candidate gene (CGPPB6189, accession number AJ876189) was also polymorphic in peach varieties with five SNPs and encodes a sucrose synthase, a central enzyme in the metabolic interplay of sucrose, hexoses, and starch synthesis. This gene co-localizes with three QTLs (glucose, fructose, and sucrose) mapped in linkage group 7 of the *Prunus* map in an advanced backcross between *Prunus persica* cultivars and the wild relative species *Prunus davidiana* (Quilot et al 2004).

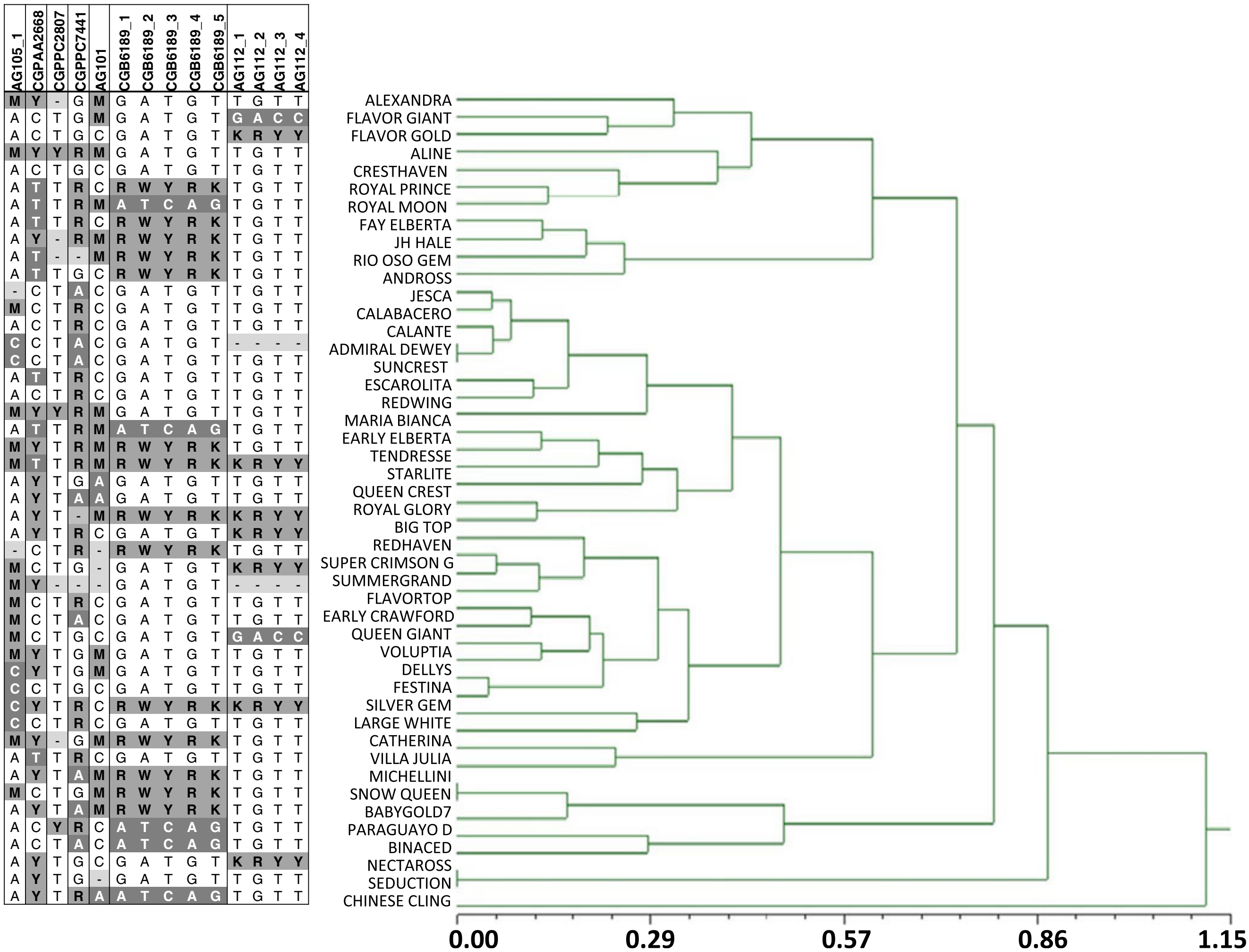


Fig. 3 SNP alignment matrix and UPGMA tree constructued with six SSRs close to the sequenced polymorphic DNA fragment

The whole sequence of the pectate lyase gene (*PpPL1*) was obtained from sequencing four fragments (including CGPAA2668). The consensus sequence contained 2,239 bp: 1,000 bp corresponded to intronic and flanking regions and 1,239 bp to coding DNA distributed in four exons. In total, the whole fragment contained two SNPs (i.e., 1 SNP every 1,119 bp), one in non-coding DNA and the other in coding DNA producing a synonymous replacement. All mutations were linked in a whole haplotype, observing two haplotypes and three genotypes (the two homozygous

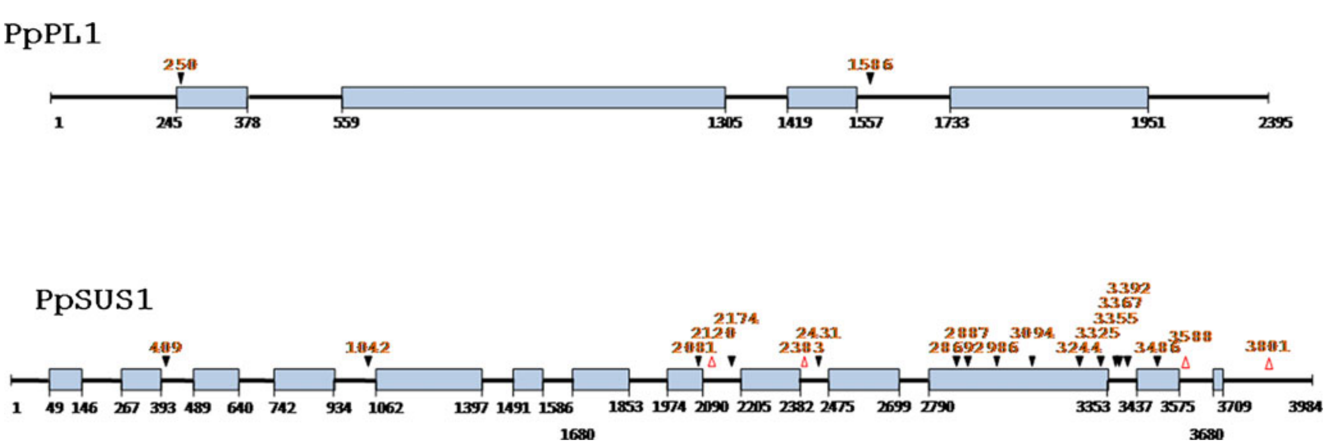


Fig. 4 Scheme of *PpPL1* and *PpSUS* genes. *Grey boxes* represent exons. *Full triangles* represent SNPs and *empty triangles* indels

plus the heterozygous). The most common allele had a frequency of 60 %. Nucleotide diversity (π) was in the range of values observed in the seven polymorphic fragments (0.00043) while θ_w was much lower (0.00017).

The whole sequence encoding the sucrose synthase (*PpSUS*) was obtained from sequencing seven fragments, producing a consensus sequence of 3,984 bp (3,941 bp excluding gaps): 1,569 bp corresponded to intronic and flanking regions and 2,415 to coding DNA distributed in 13 exons. In total, the whole fragment contained 15 SNPs and four indels (one SNP every 263 bp and one indel every 985 bp). Seven of the SNPs occurred in six introns and eight in three exons, all were synonymous changes except one which produced the replacement from a lysine (the most frequent) to an asparagine. This replacement is likely to have a limited effect on the *PpSUS* enzyme activity due to the similar physicochemical properties of these two amino acids. All indels occurred in non-coding regions, three in introns and one in the 3′UTR. All SNPs were linked in a whole haplotype. Three of the indels were linked to the SNPs while the fourth, of 19 bp, was only observed in the Spanish landrace "Jesca". No recombination was detected in the whole fragment, with three haplotypes and four genotypes observed. After removing indels, two haplotypes and three genotypes were observed. Nucleotide diversity (π) and θ_w values were within the range of values observed in the seven polymorphic fragments (π = 0.00146, θ_w=0.00074) (Fig. 1). The most abundant SNP haplotype had a frequency of 74.5 %, with 28 of the varieties homozygous for the most frequent allele (haplotype) and 14 for the less frequent.

For both genes, Tajima's *D* values, indicative of selection, were significantly higher than zero (2.301 and 2.696 for the pectate lyase and the sucrose synthase, respectively; $p \leq 0.05$), indicating an excess of alleles at intermediate frequency, possibly due to balancing selection.

Here, we observed relatively low SNP polymorphism in peach, consistent with the low variability previously described in the species. One of the two genes sequenced had an SNP density higher than that observed at genome-wide level and a possible pattern of selection was observed in both. This, together with their map position and putative gene function make them good candidates for affecting the phenotype. To provide additional evidence on the causal effects of these genes on the peach fruit phenotype, a larger sample of cultivars should be genotyped and phenotyped for different components of fruit firmness and sugar content to detect association between these two genes and the phenotype in which they could be involved.

Acknowledgments This research was funded in part by Projects AGL2009-07305/AGR and by the Consolider-Ingenio 2010 Program (CSD2007-00036), both from the Spanish Ministry of Science and Innovation, and by the ISAFRUIT Integrated Project. The ISAFRUIT Project was funded by the European Commission under Thematic Priority 5—Food Quality and Safety of the 6th Framework Programme of RTD (Contract No. FP6-FOOD-CT-2006-016279).

Disclaimer Opinions expressed in this publication may not be regarded as stating an official position of the European Commission.

References

Aranzana MJ, Carbó J, Arús P (2003a) Microsatellite variability in peach [*Prunus persica* (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. Theor Appl Genet 106:1341–1352

Aranzana MJ, Carbó J, Arús P (2003b) Using amplified fragment-length polymorphisms (AFLPs) to identify peach cultivars. J Amer Soc Hort Sci 128:672–677

Aranzana M, Abbassi E-K, Howad W, Arús P (2010) Genetic variation, population structure and linkage disequilibrium in peach commercial varieties. BMC Genet 11:69

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol J 7:347–354

Byrne DH (1990) Isozyme variability in four diploid stone fruits compared with other woody perennial plants. J Hered 81:68–71

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski A (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Deng H-W, Chen W-M, Recker RR (2001) Population admixture: detection by Hardy-Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. Genetics 157:885–897

Dirlewanger E, Cosson P, Tavaud M, Aranzana MJ, Poizat C, Zanetto A, Arús P, Laigret F (2002) Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry *(Prunus avium* L.). Theor Appl Genet 105:127–138

Dirlewanger E, Graziano E, Joobeur T, Garriga-Calderé F, Cosson P, Howad W, Arús P (2004) Comparative mapping and marker-assisted selection in Rosaceae fruit crops. Proc Natl Acad Sci USA 101:9891–9896

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE 2:e1367

Howad W, Yamamoto T, Dirlewanger E, Testolin R, Cosson P, Cipriani G, Monforte AJ, Georgi L, Abbott AG, Arús P (2005) Mapping with a few plants: using selective mapping for microsatellite saturation of the *Prunus* reference map. Genetics 171:1305–1309

Illa E, Eduardo I, Audergon JM, Barale F, Dirlewanger E, Gao ZS, Moing A, Lambert P, Le Dantec L, Li XW, Poëssel JL, Pozzi C, Rossini L, Vecchietti A, Arús P, Howad W (2011) Saturating the *Prunus* (stone fruits) genome with candidate genes for fruit quality. Molec Breed 28:667–682

Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, Main D (2008) GDR (Genome database for Rosaceae): integrated webdatabase for Rosaceae genomics and genetics data. Nucleic Acids Res 36 (Database Issue): D1034-D1040

Keller I, Bensasson D, Nichols RA (2007) Transition–transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 3(2):e22

Labate J, Robertson L, Wu F, Tanksley S, Baldo A (2009) EST, COSII, and arbitrary gene markers give similar estimates of nucleotide diversity in cultivated tomato (*Solanum lycopersicum* L.). Theor Appl Genet 118:1005–1014

Laval G, SanCristobal M, Chevalet C (2002) Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genet Sel Evol 34:481–507

Lewis PO, Zaykin D (2001) Genetic data analysis: computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from http://alleyn.eeb.uconn.edu/gda/.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452

Lijavetzky D, Cabezas J, Ibanez A, Rodriguez V, Martinez-Zapater J (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics 8:424

Marin-Rodriguez MC, Orchard J, Seymour GB (2002) Pectate lyases, cell wall degradation and fruit softening. J Exp Bot 53:2115–2119

Micheletti D, Troggio M, Zharkikh A, Costa F, Malnoy M, Velasco R, Salvi S (2011) Genetic diversity of the genus *Malus* and implications for linkage mapping with SNPs. Tree Genet Genomes 7:857–868

Miller PJ, Parfitt DE, Weinbaum SA (1989) Outcrossing in peach. HortScience 24:359–360

Mnejja M, Garcia-Mas J, Audergon J-M, Arús P (2010) *Prunus* microsatellite marker transferability across rosaceous crops. Tree Genet Genomes

Morales M, Roig E, Monforte A, Arús P, García-Mars J (2004) Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). Genome 47:352–360

Nei M (1987) Molecular evolutionary genetics. Columbia Univ, Press, New York

Quilot B, Wu BH, Kervella J, Génard M, Foulongne M, Moreau K (2004) QTL analysis of quality traits in an advanced backcross between *Prunus persica* cultivars and the wild relative species *P. davidiana*.TAG. Theor Appl Genet 109:884–897

Rohlf FJ (1994) NTSYS-pc. 2.02 edn. Exeter Softwares, Setauket, New York

Rozen S, Skaletsky H (2000) Primer3 on the www for general users and for biologist programmers. Methods Mol Biol 132:365–386

Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, Zyprian E, Toepfer R, Stella Grando M, Velasco R (2004) Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera* L.), as revealed by single nucleotide polymorphisms. Mol Breed 14:385–395

Scorza R, Mehlenbacher SA, Lightner GW (1985) Inbreeding and coancestry of freestone peach cultivars of the Eastern United States and implications for peach germplasm improvement. J Amer Soc Hort Sci 110:547–552

Simko I, Haynes KG, Jones RW (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173:2237–2245

Sosinski B, Gannavarapu M, Hager LD, Beck LE, King GJ, Ryder CD, Rajapakse S, Baird WV, Ballard RE, Abbott AG (2000) Characterization of microsatellite markers in peach *Prunus persica* (L.) Batsch. Theor Appl Genet 101:421–428

Strandberg AKK, Salter L (2004) A comparison of methods for estimating the transi-tion:transversion ratio from DNA sequences. Mol Phylogenet Evol 32:495–503

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Testolin R, Marrazzo T, Cipriani G, Quarta R, Verde I, Dettori MT, Pancaldi M, Sansavini S (2000) Microsatellite DNA in peach (*Prunus persica* L. Batsch) and its use in fingerprinting and testing the genetic origin of cultivars. Genome 43:512–520

Till BJ, Colbert T, Codomo C, Enns L, Johnson J, Reynolds SH, Henikoff JG, Greene EA, Steine MN, Comai L, Henikoff S (2006) High-throughput TILLING for Arabidopsis. Methods Mol Biol 323:127–135

Tiret L, Cambien F (1995) Departure from Hardy–Weinberg equilibrium should be systematically tested in studies of association between genetic markers and disease. Circulation 92:3364–3365

Viruel MA, Messeguer R, de Vicente MC, Garcia-Mas J, Puigdomènech P, Vargas F, Arús P (1995) A linkage map with RFLP and isozyme markers for almond. Theor Appl Genet 91:964–971

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519

Wünsch A, Carrera M, Hormaza JI (2006) Molecular characterization of local Spanish peach [*Prunus persica* (L.) Batsch] germplasm. Genet Resour Crop Ev 53:925–932

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134